

UNITED STATES PATENT APPLICATION

For

VIDEO PROCESSING

Inventors:

Graham Alexander Thomas
Peter Brightwell
Oliver Grau

Prepared by:

BLAKELY SOKOLOFF TAYLOR & ZAFMAN LLP
12400 Wilshire Boulevard
Los Angeles, CA 90025-1026
(408) 947-8200

Attorney's Docket No.: 7006P001

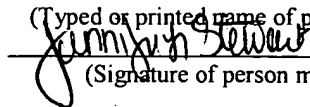
"Express Mail" mailing label number EV410138129US

Date of Deposit March 12, 2004

I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee is addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Jennifer Stewart

(Typed or printed name of person mailing paper or fee)



(Signature of person mailing paper or fee)

VIDEO PROCESSING

PRIORITY INFORMATION

[0001] This application claims benefit and priority to United Kingdom Application No. 0305926.8 Filed 14 March 2003 entitled "Video Processing", the contents of which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

[0002] This invention relates to video processing, and more specifically to virtual image production. The present invention may be used in a number of different areas of video and image production, but is particularly applicable in the field of television sports coverage.

[0003] The use of virtual reality techniques is becoming increasingly common in television and video production, however application in sports coverage is at present relatively limited. Replays, slow motion and detailed analysis of sports events are popular, and there is a growing desire to be able to provide computer enhanced images and sequences for these purposes. A wide variety of virtual techniques have been proposed in the field of video and television production.

[0004] Examples of prior art techniques in the field of sports coverage include the Epsis system produced by Symah Vision, which is regularly used to provide tied-to-pitch logos, scores, distance lines, etc. for football, rugby, and other sports. This system is limited however to relatively simple graphics, and works with a camera at a fixed position. It would be desirable to provide more sophisticated image and video manipulations of live action events such as sports coverage.

[0005] An example of a desirable effect would be to provide the viewer with a specific view of a scene, such as a view along a finish line or an offside line. In the case of a static finish line the solution of arranging a camera looking along that line is trivial. Where desirable views cannot be predetermined (such as an offside line) a number of possible approaches have been proposed.

[0006] Arranging a multitude of cameras along the side of the pitch, so that one camera will give approximately the desired view is one such proposal. EyeVision from Princeton Video International [<http://www.pvi-inc.com/>] uses this approach with

cameras typically arranged in a circle or arc. However the large number of cameras required to achieve a sufficiently precise view makes this solution too costly or impractical for many events.

[0007] A moving camera is an alternative proposal. A number of systems exist for cameras on rails and wires (e.g. [www.aerialcamerasystems.com]), however it cannot be guaranteed that the camera will be in the right place in the right time to produce the desired image, and the producer cannot change his/her mind after the event.

[0008] Another approach is provided by Orad's Virtual Replay system [www.orad.co.il]. This uses image-processing based techniques including white-line matching to determine the camera parameters and player tracking, and renders a complete virtual image of the scene including the pitch, stadium and players as 3D graphics. This is an expensive solution, and quite slow in use. A particular disadvantage of this system for sports coverage is that the virtual players may be considered to look too generic, and that a large amount of detail in a scene may be lost when scenes are rendered. It is recognised, however, that the intention of this system is not to provide a realistic image and there may be some attractions to the "computer game" image generated.

[0009] A further approach is disclosed in US-A-4,956,706. This provides a method of manipulating a camera image to form a view of a scene from a greater elevation than that of the camera image. This is done by calculating a planar model of the scene, and applying the camera image to this model by using a stretch transformation. In order to compensate for items having any significant height, the planar model can be locally deformed by defining a deformation vector and a point of action on the planar model. This method is intended to be used with generally planar scenes where a low level of detail is required, for example an overhead view of a golf course, and hence is not intrinsically applicable to providing a virtual viewpoint of a generalised 3-D scene, which would require the entire planar model to be substantially deformed in a very complex manner. It is not disclosed how to determine which picture areas require local deformation, which apparently requires manual identification. This would not be practicable for a dynamically changing scene.

[0010] It is an object of the present invention to provide an improved method of creating a view of a real scene from a selected viewpoint. The term viewpoint as used herein may include both a position or direction from which a view is obtained and a zoom or magnification factor or field of view parameter.

BRIEF SUMMARY OF THE INVENTION

[0011] Accordingly, in a first aspect the invention provides a method for generating a desired view of a real scene from a selected desired viewpoint, said method comprising:

- obtaining at least one real scene image from one or more cameras, the or each camera having a respective real viewpoint;

- identifying selected objects in said at least one real scene image;

- determining estimates of the positions of the selected objects;

- selecting a desired viewpoint;

- based on the relationship of the selected desired viewpoint to the or each real viewpoint, determining positions of the selected objects in said desired view of the scene and rendering a view of the scene from the selected desired viewpoint wherein at least some selected objects are rendered using image data from at least one real scene source image.

[0012] In this way, real image data is used to render selected objects (e.g. players or groups of players in a rugby scrum for example) and the impression given is of a much more realistic view. The source image may be a preceding image in a sequence of images, but will normally be a co-timed image. Other portions, or the remainder of the view can be rendered from alternative data. This method allows the most important parts (eg. players or the ball) of the virtual view from the desired viewpoint to be accurately rendered by using time varying, current image data, while less important parts (eg. pitch and crowd) can be rendered less accurately using less critical data, which may be generic and/or time invariant.

[0013] Optionally a portion of the image, optionally the background portion is generated without accurate transformation of real image data, for example by using known virtual rendering techniques. For example a grass field or other area may be generated by synthesising an appropriate texture and field markings. However, elements of texture or colour for use in the synthesis may be derived from real image data, for example by obtaining a texture sample. Using the example of a football stadium, the pitch and crowd can be rendered from a computer model describing the geometry of the stadium, with texture taken from pre-recorded footage of the stadium, possibly when empty or from a previous game, since it is not important for this data to be co-timed in the rendered virtual view.

[0014] Optionally all selected objects are rendered using real image data but the technique may be applied to designate two categories of selected objects, a first category (e.g. key players) to be rendered using real image data, a second category (e.g. players further from key action) to be rendered using virtual representations.

[0015] The step of identifying a selected image object is optionally performed using a real scene image by a keying process, and more optionally by a chroma keying process, which can be used to good effect to separate images of sportsmen from a background of a grass surface for example. Alternatively, where a sequence of real scene images are obtained from a camera, difference keying may be used. In certain situations it may be desirable to allow for a degree of user intervention in the keying process, or even to allow a user to indicate approximately or by more accurate tracing around some or all selected objects in a real scene image. Depth keying is a further possibility for some applications.

[0016] The position of objects in a scene can be calculated from a single camera image of that scene and a constraint, or from multiple camera images as explained below. In this way an estimate of the 3-D (or 2-D and a constrained third dimension) position of the selected objects can be derived and used in producing the rendered view from the desired viewpoint.

[0017] Optionally selected objects in the desired view are rendered as projections of real images of those objects obtained from said real scene image, optionally by transforming real image data based on the relationship of the real viewpoint of the camera from which the image is taken and the selected desired viewpoint. In a simple embodiment, real images of the selected objects are obtained and used as flat models oriented perpendicular to the optical axis of the real camera. These models can then be rendered from the point of view of the selected viewpoint by projection. This simple approach has been found to produce surprisingly good results, particularly when the selected viewpoint and the real camera viewpoint differ in angle by less than approximately 30 degrees.

[0018] In some cases, beneficial results may be achieved by obtaining images of selected objects, and allowing the images to be rotated when modelling the objects. The objects can be rendered from a selected viewpoint by rotating the images, either partially up to a defined limit or up to an amount which is a function of the angle between the real and desired viewpoint or to be perpendicular to the optical axis of the selected viewpoint. In this way the resolution of the images is not reduced, which may

be advantageous where the image is already of low resolution. In some situations it may be desirable to render objects with the image 'models' at different angles of rotation. The angle of rotation of an image may be determined by a user, may be determined automatically based on, for example, the object's direction of movement, or may be determined by a combination of these factors. A potential disadvantage of this approach is that it may produce artefacts in a video sequence of virtual images in which the selected viewpoint moves.

[0019] A further enhancement in image rendering is to model selected objects as images of those objects mapped onto approximate 3D surfaces, for example a rounded object rather than a flat panel. These models can then be rendered from selected viewpoints. This provides a more realistic virtual image, and may allow an object to be more satisfactorily rendered from a wider range of selected viewpoints for a particular given real scene image.

[0020] Optionally the 3D surface onto which an image is mapped is derived from the outline of that image. Techniques for producing such a 3D surface are known, and typically make some assumptions about the curvature of bodies. Shape from silhouette is an example of a technique which has been developed to provide a rough 3D surface from multiple 2D images of an actor, and an improved technique is disclosed in our earlier UK patent application No. GB 0302561.6, the entire disclosure of which is incorporated herein by reference. Where simplifying assumptions about the selected objects can be made it is possible to produce an approximate 3D surface onto which an image can be mapped from a single 2D image.

[0021] An additional aspect of the invention provides apparatus for generating a desired view of a real scene from a selected desired viewpoint, comprising:

- means for obtaining at least one real scene image from one or more cameras, the or each camera having a respective real viewpoint;

- means for identifying selected objects in said at least one real scene image;

- means for determining estimates of the positions of the selected objects;

- means for selecting a desired viewpoint; and

- based on the relationship of the selected desired viewpoint to the or each real viewpoint, means for determining positions of the selected objects in said desired view of the scene and rendering a view of the scene from the selected desired viewpoint wherein at least some selected objects are rendered using image data from at least one real scene source image.

[0022] It is of course possible that in object may be partially obscured in a real scene image. In order to render such object in the virtual image, it is possible digitally to synthesise part of the real image of that object. This is optionally achieved by interpolation between successive images in a sequence. This approach may not be appropriate however when an image at a certain instant in time is required. An alternative approach is to match missing image data with data from another part of the same real scene image. It will be appreciated that conventional image prediction and correction techniques can be applied for this novel purpose.

[0023] One particularly preferred embodiment of the invention includes providing more than one real camera to provide a set of different real scene images, each real scene image corresponding to a different viewpoint. An immediate advantage of this embodiment is that a wider range of possible viewpoints may be selected for which there is a real scene image at a sufficiently close angle to produce acceptable renderings of objects. Another important advantage is that when an object is obscured or partially obscured in one real scene image, it may be possible to use an image of that object from another real viewpoint in which the object is not obscured, or at least in which the same part of the object is not obscured. Rendering may include selecting a preferred image source for each selected object.

[0024] In a simple example of an embodiment having a plurality of real cameras, selected objects are rendered in the virtual image using image data from the real scene image whose corresponding viewpoint is closest to the selected viewpoint. This example can be extended by using image data from other real scene images for rendering a selected object when the 'closest' real scene image shows that object either partially or totally obscured. An iterative selection process for selecting an appropriate real scene image to render an object may be employed based on a number of criteria, such as the difference in angle of the selected view from the real view, and the coverage of the selected object. Where no appropriate image for a selected object can be found based on selected criteria, it may be desirable not to include that image in the virtual view. Alternatively a weighting factor could be calculated for an object based on selected criteria, and the representation of that object could be faded in and out of the virtual image according to that weighting factor. This could be implemented using an alpha signal for pixel transparency.

[0025] In a more advanced example selected objects are rendered in the desired view using image data from two or more of a set of real scene images. A cross fade

between two real viewpoints could be used for a desired view from a selected viewpoint between the two real viewpoints, and this can be weighted according to the ratio of distance between the two real viewpoints. This might be used to particularly good effect for producing a video sequence of views from different selected viewpoints. A more complex alternative would be to use a form of motion compensated interpolation, such as FloMo, produced by Snell & Wilcox. This would be unsuitable for live use however, since extensive post processing is required.

[0026] The use of multiple real cameras can be advantageously exploited in embodiments where selected objects can be modelled as real images mapped onto a 3D surface. A suitable 3D surface can be created from the intersections of generalised cones of the outline of a selected object viewed from different real viewpoints. A generalised cone is the union of visual rays from all silhouette points of a particular image. This intersection gives an approximation of the real object shape and is called the visual hull. Several algorithms have been published for the computation of the visual hull, for example: W. Martin and J. K. Aggarwal, "Volumetric descriptions of objects from multiple views," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 5, no. 2, pp. 150-158, March 1983.

[0027] The use of multiple real cameras to monitor and locate objects in a scene, and to provide image data of those objects for rendering purposes may be provided independently in one aspect of the invention. This aspect of the invention provides a method of monitoring a scene for virtual image generation, said method comprising:

- obtaining a set of real scene images from a plurality of cameras having mutually different viewpoints;

- using image data from at least a first of said real scene images to derive the location of a selected object in the scene; and

- using image data from at least a second of said real scene images to render a virtual image of said selected object.

[0028] A related aspect of the invention provides apparatus for monitoring a scene for virtual image generation, said apparatus comprising:

- means for obtaining a set of real scene images from a plurality of cameras having mutually different viewpoints;

- means for using image data from at least a first of said real scene images to derive the position of a selected object in the scene; and

means for using image data from at least a second of said real scene images to render a virtual image of said selected object.

[0029] Optionally first and second subsets of images are used respectively for location and rendering but equally, all images may be used. Each subset, particularly the second subset, may comprise only images from a single camera. The subsets may overlap but are optionally non-identical. Optionally, the first subset of images includes at least one image from a camera having an elevated viewpoint of the scene, and the second subset includes at least one image from a camera having a low-level viewpoint of the scene. One advantage of this arrangement is that objects are less likely to be obscured in a real image of a scene obtained from an elevated viewpoint. Although images from elevated viewpoints may not be particularly useful for rendering purposes when it is desired to generate a virtual image from a low level viewpoint (as is often the case), such images are still useful for determining the 3D position of objects in the scene. It is desirable to be able to track selected objects in one or more sequences of real images, and this can often be performed more easily using images from elevated viewpoints for the reasons given above. It has been found that it is not necessary to provide a high level camera corresponding to each low level camera, and that in fact, the total number of cameras can be reduced by providing high and low level cameras, at mutually different lateral orientations around a scene. This solution provides a good working compromise.

[0030] Although it has been shown that providing more than one real camera can provide a number of benefits, there is the potential disadvantage that an equivalent number of camera operators may be required. This problem can be overcome in an embodiment of the invention wherein one or more cameras are slave cameras. Slave cameras can be operated automatically based on camera parameters (eg. pan, tilt, zoom and focus) from one or more other cameras to which they are linked. One preferable set up automatically controls one or more slave cameras to point towards the average centre of other real cameras, and the focus may be set, for example, at a certain height above the ground or pitch in the case of a sports application. It may be necessary to override the automatic control, or at least to modify the control algorithm in certain situations, for example when one or more controlling cameras is pointing in an unhelpful direction.

[0031] In a further aspect of the invention therefore, there is provided a method of controlling a slave camera based on the parameters of at least one other camera, said method comprising:

adjusting the parameters of said slave camera to point and focus at a desired point based on the camera parameters of at least one of said other cameras.

[0032] A still further aspect of the invention provides apparatus for controlling a slave camera based on the parameters of at least one other camera, said apparatus comprising:

means for adjusting the parameters of said slave camera to point and focus at a desired point based on the camera parameters of at least one of said other cameras.

[0033] This is an advantageous method of obtaining a number of images of a scene from different cameras, without requiring a corresponding number of camera operators. Automatically controlling the focus of said slave cameras results in images which can be used immediately and are therefore more useful eg. in a quick camera switch. It is preferable therefore, that all of the pan, tilt, zoom and focus parameters of the slave camera are controlled.

[0034] As mentioned already, it is desirable to be able to track selected objects in one or more sequences of real scene images. By tracking an object over a period of time (over a number of images) and also determining an estimate of its position at each defined instant of time, it is possible to produce a path or trajectory of that object in space. This path can usefully be displayed against a background of the scene in an analysis display, which can be provided from substantially any virtual viewpoint, even where real image data cannot be reliably rendered. In addition, by tracking objects statistics, such as instantaneous velocity and distance travelled can be derived. In order to reduce the demands on the operator it is preferable that this tracking can be performed automatically. In a preferred embodiment tracking is performed by obtaining a silhouette or outline of selected objects from a real scene image (and optionally from a real scene image from an elevated viewpoint), for example by keying, and analysing changes in shape or position of this silhouette from frame to frame. More optionally there is provided a user interface to allow an operator to view one or more real scene images, and to manually adjust the tracking of one or more selected objects. This may be performed by manually selecting the position of a tracked object on one or more images at a given time. This feature is particularly beneficial in applications where selected objects change shape and overlap, for

example where selected objects are players in a rugby match. Additionally, the user interface can be arranged to allow an operator to adjust the keying of a selected object in one or more real scene images.

[0035] In a further aspect of the invention therefore, there is provided apparatus for tracking selected objects in a scene comprising:

one or more cameras arranged to obtain one or more real scene images;

image processing means for identifying said selected objects in said one or more real scene images;

means for providing an estimate of the three-dimensional spatial position of said one or more selected objects based on their position in the one or more real scene images; and

a user interface adapted to allow an operator to view said estimate of the position of selected objects in a real scene image, said user interface including input means to allow an operator to modify said estimate.

[0036] This novel apparatus reduces the demands on an operator by providing an automatic estimate of position, while at the same time allowing a degree of human intervention in cases where the estimate is incorrect, or when no estimate can be produced. A variable degree of control may be provided to the operator.

[0037] It is possible to provide an automatic estimate of position using a single image of a scene when an estimate based on an assumption about a constraint can be made. One such assumption is that selected objects are in contact with the ground, or constrained to a reference surface. Assumptions about the size or shape of a selected object can also be used in some circumstances, for example assuming the height of a player in a sports match.

[0038] Optionally a plurality of cameras is used to obtain a plurality of real scene images, each said image corresponding to a different viewpoint. This allows a more accurate estimate of the position of objects, particularly in cases where objects are obscured from certain views.

[0039] Where multiple real cameras are used it is desirable that the user interface allows an operator to view images from more than one camera simultaneously. Optionally the user interface provides the operator with an automatic estimate of the three dimensional position of selected objects in the real scene derived from one or more real scene images, through the use of simultaneous displays. In this way an

operator may correct or adjust the automatic estimate, optionally by interaction with one of the displayed real scene images.

[0040] The user interface optionally also allows the operator to select real scene images which should be used to track and locate selected objects. In this way information from a camera pointing in a direction which is not useful for object tracking (eg. a camera pointing at the crowd in a football match) can be selectively disregarded. The same user interface may desirably be used to control the operation of slave cameras by selecting which real cameras should provide control information to a given slave camera.

[0041] In a particular embodiment of the invention used in television production of sports matches, and in particular football, it is desirable to obtain an estimate of the position of the ball in the scene. Obtaining an accurate estimate has proved to be difficult in the past, on account of the fact that the ball is relatively small, and is not always on the ground. The user interface may advantageously be adapted to provide an improved estimate of the ball position based on images of the ball from cameras, and operator inputs. In one embodiment the user can input the location of the ball in two or more camera images to allow an estimate of position to be determined, or an estimate of the position may be presented for user selection or refinement. In an extension of this idea, the trajectory of a ball in flight can be estimated based on user defined positions of a start point and an end point of the ball's flight, and using standard calculation techniques assuming a parabolic flight. A further improvement of this feature could take into account air resistance acting on the ball.

[0042] Another aspect of the invention provides A computer program or a computer program product for generating a desired view of a real scene from a selected desired viewpoint, which when implemented performs the steps of:

- obtaining at least one real scene image from one or more cameras, the or each camera having a respective real viewpoint;
- identifying selected objects in said at least one real scene image;
- determining estimates of the positions of the selected objects;
- selecting a desired viewpoint;
- based on the relationship of the selected desired viewpoint to the or each real viewpoint, determining positions of the selected objects in said desired view of the scene and rendering a view of the scene from the selected desired

viewpoint wherein at least some selected objects are rendered using image data from at least one real scene source image

[0043] Yet another aspect of the invention provides a computer program or a computer program product for monitoring a scene for virtual image generation which when implemented performs the steps of:

obtaining a set of real scene images from a plurality of cameras having mutually different viewpoints;

using image data from at least a first of said real scene images to derive the position of a selected object in the scene; and

using image data from at least a second of said real scene images to render a virtual image of said selected object.

[0044] Still another aspect of the invention provides a computer program or a computer program product for controlling a slave camera based on the parameters of at least one other camera, which when implemented adjusts the parameters of said slave camera to point and focus at a desired point based on the camera parameters of at least one of said other cameras.

[0045] It should be understood that features may be provided independently or in combination, and although specific examples have been described, alternative embodiments are intended as falling within the scope of the invention. It is intended that this application extends to apparatus for performing methods according to the invention, and vice versa and that preferred features of methods according to the invention apply similarly to apparatus according to the invention and vice versa. Method or apparatus features described herein also apply to embodiments of the invention comprising computer programs and computer program products.

BRIEF DESCRIPTION OF THE DRAWINGS

[0046] Various embodiments of the invention will now be described by way of example with reference to the accompanying drawings in which:

[0047] Figures 1a and 1b show methods of rendering a 2D image obtained from a real camera from the point of view of a virtual camera.

[0048] Figures 2a and 2b show an alternative method of rendering a 2D image.

[0049] Figures 3a and 3b show an example of an object being obscured from a viewpoint.

[0050] Figure 4 illustrates multiple cameras being used to allow images from a range of desired positions to be rendered.

[0051] Figure 5 illustrates a multiple camera approach used in conjunction with the rendering technique of Figure 2

[0052] Figure 6 shows a camera arrangement suitable for a football game.

[0053] Figures 7a and 7b illustrate one possible source of error in a camera tracking and positioning system.

[0054] Figure 8 shows an example of a visual hull produced for a selected object.

[0055] Figures 9 and 10 are examples of possible screen outputs for one embodiment of a user interface according to an aspect of the invention.

[0056] Figure 11 is a schematic illustration of a system according to one embodiment of the present invention

DETAILED DESCRIPTION OF THE INVENTION

[0057] It can be seen in Figure 1a that using a single real camera 102 we can model a selected object 104 most simply as a 2-D plane 106 at right angles to the real camera axis 108. The images from the real camera are rendered as a flat texture from the position of the virtual camera 110. An observer at the virtual view point sees the virtual object as a “cardboard cut-out”. This approach works reasonably well when the difference between the real and virtual camera angles is up to about 30 degrees, beyond which the distortion becomes too apparent.

[0058] A variation of the 2-D approach is illustrated in Figure 1b, in which the planes modelling selected objects are rotated to a suitable angle 107. In some situations this may give a better virtual view, for example where the angle of view of the main camera is relatively narrow (otherwise the 2-D image will not have enough horizontal resolution), and the 2-D image is approximately perpendicular to the virtual camera 110.

[0059] A “2½-D” approach is illustrated in Figures 2a and 2b. A 2-D image 202 of an object 203 is obtained from a real camera 204 as shown in Figure 2a. Image 202 is then mapped onto a 3-D curved surface 206 as shown in Figure 2b. This 3-D surface model is then rendered from the position of a virtual camera 208.

[0060] The single camera approach will often be limited where one object obscures another. This is shown in Figure 3a, where object 302 cannot be rendered properly from many virtual camera angles based on the 2-D image 304 obtained from real camera 306. For games such as fifteen-a-side rugby this will be the case for a significant proportion of the time for typical camera angles. A higher camera position will reduce the amount of overlap, but this will increase the distortion of the rendered players, and such a position may not be available. Of course the situation shown in Figure 3b is perfectly acceptable, and the rendered view from virtual camera 308 will show object 310 partially obscured by object 312.

[0061] It may be possible to synthesise missing object image information by using scene images from preceding or following frames. At its simplest, this would involve simply displacing the 2-D or 2½-D textures from the previous frame to match the current position of the object. However, this should not be used where it is important to have an accurate representation of the scene, for instance to show a controversial offside decision. Alternatively motion-compensated prediction could be used on the

input video to generate the missing information. This is only likely to work reasonably when the player has been obscured for a few frames. A possibly better approach may be to try to match the missing information to something similar in another part of the frame. Unlike conventional motion estimation techniques such as block matching, the match is not assumed to be near the missing information. So a missing portion of a player's arm, for example, might be replaced by a similar-looking portion of someone else's arm. It has been proposed to use this approach with a method called "long-range correlation" to give impressive results for image restoration and error concealment. For matching large areas a hierarchical matching system could be used to reduce the computational requirements. This algorithm assumes that the missing area is to be matched with an area the same size and shape. It may also be possible to match with a different sized area using techniques suitable for fractal image coding.

[0062] Long-range correlation or fractal matching methods could be extended to search in other frames if necessary. Alternatively, in the example of a football game for example, a match could be performed against a "library" of player images which could be prepared before the game, or built up as the game progresses.

[0063] Even if one or more of the above methods are used to reconstruct the obscured parts, it is still necessary to know which parts are missing. This could be performed using segmentation methods, by inter-frame differences, or by some combination, but it is likely to be difficult in some cases, especially when two overlapping objects have a similar appearance, and therefore it is desirable to provide some user intervention.

[0064] Figure 4 shows one possible multi-camera arrangement that would be suitable for a football match rigged with a camera 402 on the centre line and one on each of the 18-yard lines (404 & 406). Each of players 410, 412 and 414 can be seen unobscured from at least one real camera. Player 410 can be rendered from a reasonable angle by a virtual camera at any point along path 416, by using the 2-D technique described above from the most appropriate camera.

[0065] For a view from virtual camera 420, player 410 is rendered using the video from camera 402 and for a view from virtual camera 422, player 410 is rendered using the video from camera 404. At some point between virtual camera positions 420 and 422 there will be a noticeable switching effect. Alternatively a cross-fade between the two camera views could be used although is arguably less acceptable to the viewer.

“Motion”-compensated interpolation could be employed to interpolate between the views from two positions, although this has typically required a lot of hand-crafting in the post processing so is not suitable for live use.

[0066] However, even with three cameras, there are still problems. It will not be possible to obtain a good view of player 412 from a virtual camera on the left hand side of path 416, because only camera 406 provides a full real view. In views from virtual cameras 420 and 422, player 412 is obscured by player 410, but at in a view from virtual camera 424 player 412 can be seen. To prevent unwanted distortions, players can be “faded out” as the angle of the virtual camera becomes too great.

[0067] Figure 5 illustrates a multiple camera set up using the “2½-D” approach. As described previously, real image segments (eg. 502, 504) are mapped onto 3D surfaces as textures. More than one real image segment derived from more than one real camera can be mapped onto a single 3D surface representing a selected object or player. This is the case for player 510, where image segments 506, 507 & 508 are derived from cameras 526, 528 & 530 respectively. In Figure 5 the virtual view of player 512 might just be acceptable in a view from virtual camera 524. However in general, more than three cameras are likely to be required to provide a good range of reliable virtual camera angles when there are many players on the pitch.

[0068] Figure 6 shows seven cameras used at a football match. Most of the 23 players (including referee) can be viewed from most virtual angles (on one side of the pitch), but there are still some exceptions. For instance the player 602 cannot be fully viewed from the bottom left or left. High camera positions will reduce this effect, and are more suitable for player tracking, but will increase the distortion when rendering a virtual camera view from a low angle. In practice it would be best to have a combination of high and low camera angles. In Figure 6 cameras 610, 614, 618 & 622 would typically be mounted at low-level, while cameras 612, 616 & 620 would typically be elevated. If it proves necessary to have more real cameras available than there are camera operators, additional slave cameras could be used. The pan, tilt, zoom and focus of the slave cameras would be set automatically using the settings of the manually operated ones. Certain assumptions will need to be made, for example that the slave cameras should be pointing at the average centre of the real cameras, and focused to a point 1.5 metres above the ground at this point. It will also be necessary to detect when the manual cameras are pointing at something different, e.g. the crowd.

[0069] More cameras, especially at different heights, will also help overcome an additional problem exemplified in Figures 7a and 7b. Here it can be seen that if we just use the key information from two real cameras 702 & 704, we can interpret the scene in two different ways. To determine the correct interpretation a constraint called “photo-consistency” can be used which uses the image data (not just the key) to estimate the position of selected objects. Techniques to address photo-consistency have previously been proposed, (eg. <http://www.cs.cornell.edu/rdz/Papers/KZ-ECCV02-recon.pdf>) but are in general very computer-intensive, although it may be possible to simplify the process in cases such as Figure 7 where there are only two possibilities. Alternative methods of preventing wrong interpretations include making certain assumptions about the sizes of objects, predicting the position and orientation of objects from previous frames; or introducing a degree of manual input. Utilising an additional camera position providing images from an elevated view point makes the disambiguation process easier.

[0070] Where more than one camera is used, shape from silhouette techniques can be used to generate approximate 3D volumes for objects in images. We will consider a simple illustration in only two dimensions with two real cameras. The outline of a simple object, such as a circle, will subtend a viewing arc at each viewpoint. The edges of these two viewing arcs intersect at four points that can be joined to form a quadrilateral which is tangent to the circle on each side. In the illustration this quadrilateral shape can be used as the basis of a simple 3D surface onto which an image can be mapped. More complicated shapes, and hence 3D surfaces can be generated with a greater number of real cameras. This technique tends to produce angular shapes and surfaces, which are optionally rounded off.

[0071] Figure 8 is a schematic representation of a ‘visual hull’ constructed for an object 802 viewed from three cameras. Images of object 802 would be rendered as texture onto a shape based on the hexagon 804 bounded by the core of rays (eg. 806 & 808 for camera 3) from the three cameras as shown in Figure 8. A more realistic appearance can be achieved by rounding off the corners of the hexagon. The texture is typically generated from the real camera closest to the virtual viewpoint.

[0072] In an example of the invention used in sports coverage, it is desirable to track players automatically, to reduce the demands on the operator. This can be done using the key signal to generate a silhouette and attempting to determine how this changes from frame to frame. However in general player tracking can be difficult, as

players change shape and overlap. This is especially true for sports such as rugby, where there are more players and there are frequent tackles, scrums, and rucks, etc. As the player tracking may fail from time to time, it is desirable to provide a user interface to allow an operator quickly to correct things.

[0073] One possible such user interface is exemplified in Figures 9 and 10. The players that the system is tracking and have been previously identified are shown with a white ellipse 902 and the name of the player 904. A yellow ellipse 906 shows players that are being tracked, but have not yet been identified. The operator can click on any player and set the current name. The interface also shows how well the keying works by colouring the player silhouettes magenta. If the operator considers the keying is incorrect, he/she can manually define the edges of the player e.g. by opening a close-up window using the user interface, e.g. by editing a “lasso selection” around the player.

[0074] Where the tracking fails, typically because of unresolvable overlaps, a red ellipse 1002 is drawn around the unknown areas, as shown in Figure 10. If appropriate, the operator can then manually draw around each player, otherwise as the players come out of overlap, the operator can wait for the red ellipse to separate into multiple yellow ellipses and identify each. If the operator chooses not to separate the players manually, they could still be rendered as a single texture. In situations where the virtual camera does not move too far this may provide an acceptable result.

[0075] In a multiple camera system, the interface could include such a display from each camera, together with a virtual display from above. This would enable the operator to quickly see how well the tracking system is doing, and use the most appropriate view to identify players. Clicking on, or moving the mouse over, a player in one view should highlight the player in all views, and this should make it obvious to the operator where the wrong estimate of position had been made. The user interface could also allow the operator to tell the system to ignore the output from certain cameras, e.g. if they are pointing at the crowd. This information could also be used to tell a system controlling slave cameras to ignore the parameters of irrelevant real cameras.

[0076] Figure 11 shows a plurality of cameras 1102 arranged to provide images of a scene 1104 (here a football pitch). The images are fed to a multiplexer 1106, and the to a central processing unit 1108. Also connected to the CPU are an image segmenter / keyer 1110, position estimation means 1112 and image rendering means 1114. A

user interface 1116 is provided which may pass data to or from the CPU. The user interface includes multiple screens, and input devices such as a keyboard 1120 and a mouse 1122. In some embodiments the user interface may comprise a PC. An image output 1124 is produced for broadcast or recording.